Probabilistic precipitation type forecasting based on GEFS ensemble forecasts of vertical temperature profiles

Michael Scheuerer*, Scott Gregory

University of Colorado, Cooperative Institute for Research in Environmental Sciences, and NOAA/ESRL, Physical Sciences Division

Thomas M. Hamill

NOAA/ESRL, Physical Sciences Division

Phillip E. Shafer

NOAA/NWS/OST, Meteorological Development Laboratory

Corresponding author address: Michael Scheuerer, University of Colorado, Cooperative Institute for Research in Environmental Sciences, and NOAA/ESRL, Physical Sciences Division 325 Broadway, R/PSD1, Boulder, CO 80305.

E-mail: michael.scheuerer@noaa.gov

ABSTRACT

A Bayesian classification method for probabilistic forecasts of precipitation type is presented. The method considers the vertical wetbulb temperature profiles associated with each precipitation type, transforms them into their principal components, and models each of these principal components by a skew normal distribution. A variance inflation technique is used to de-emphasize the impact of principal components corresponding to smaller eigenvalues, and Bayes' theorem finally yields probability forecasts for each precipitation type based on predicted wetbulb temperature profiles. Our approach is demonstrated with reforecast data from the Global Ensemble Forecast System (GEFS) and observations at 551 METAR sites, using either the full ensemble or the control run only. In both cases, reliable probability forecasts for precipitation type being either rain, snow, ice pellets, freezing rain, or freezing drizzle are obtained. Compared to the Model Output Statistics (MOS) approach presently used by the National Weather Service, the skill of the proposed method is comparable for rain and snow and significantly better for the freezing precipitation types.

1. Introduction

Some forms of winter precipitation can have a substantial impact on air and ground transportation, and reliable predictions of them can help limit associated safety hazards and disruptions of travel and commerce (Stewart et al. 2015, and references therein). Among several factors that control the precipitation type at the surface, the vertical profile of wetbulb temperature T_w plays a key role (e.g. Bourgouin 2000), and a number of algorithms have been devised which determine the precipitation type based on the T_w profile or quantities derived from it (e.g. Ramer 1993; Baldwin et al. 1994; Bourgouin 2000; Schuur et al. 2012). A major challenge herein is the model uncertainty about the T_w profile on the forecast day; while the above mentioned algorithms still show good skill in detecting snow (SN) and rain (RA), reliable distinction between ice pellets (IP) and freezing rain (FZRA) becomes increasingly difficult when this uncertainty is accounted for (Reeves et al. 2014). A recently proposed algorithm, the spectral bin classifier (Reeves et al. 2016), pushes the limits of forecast accuracy for IP and FZRA by calculating the mass fraction of liquid water for a spectrum of hydrometeors as they descend from the cloud top to the surface, thus accounting for different rapidity of melting and refreezing of smaller hydrometeors compared to larger ones. Their results still confirm the sensitivity of classification algorithms to perturbations of the T_w profile. In a forecast setting where these profiles are derived from NWP model output deviations of the true from the predicted (and interpolated) wetbulb temperature profile can be substantial, especially for longer forecast lead times. In those situations with large uncertainty it may be more useful to provide probability forecasts of each precipitation type, thus communicating the risk for precipitation to occur in the form of FZRA, say, instead of stating the most likely outcome only. Operationally, such probabilistic guidance is currently provided for the contiguous U.S. (CONUS) and Alaska by the Meteorological Development Laboratory (MDL) to support the

National Digital Guidance Database (NDGD). It is based on a model output statistics (MOS) approach that is described in Shafer (2010). This method links the probability of precipitation type (PoPT) to NWP model output of variables such as 2-m temperature, 850 hPa temperature, 1000-850 hPa thickness, 1000-500 hPa thickness, and freezing level. This approach yields conditional probabilities of freezing (IP or FZRA), frozen (SN) and liquid (RA) precipitation for forecast lead times up to 192 hours, but it does not attempt to distinguish the different freezing types. In this paper we describe an alternative method which uses (discretized) vertical wetbulb temperature profiles as a predictor, thus aiming to use more information from that profile as well as statistically modeling the forecast uncertainty. In Sec. 2 we describe the forecast and observation data used in this study, which is identical to the data used by Shafer (2015), and which thus permits a direct comparison between our approach and the operational method. Our statistical model and the methods for fitting it to the training data are detailed in Sec. 3, while a detailed evaluation of the precipitation type probabilities obtained with this model is the subject of Sec. 4. We finally discuss the scope of our method and avenues for further improvement.

2. Data used in this study

a. Observations

Adopting the setup used by Shafer (2015), our method is calibrated with and validated against weather observations at METAR (Meteorological Terminal Aviation Routine Weather Report) sites (Allen and Erickson 2001a,b). Precipitation type observations were considered for the period 1996-2013 and all months between September and May (the period 09/1996 - 05/1997 will be referred to as the '1996 cool season'), whenever precipitation was reported at the corresponding

site. Following Shafer (2015), we discarded sites where more than 50% of the precipitation type reports were missing, leading to a set of 551 stations (506 CONUS, 26 Alaska, 19 Canada).

The original precipitation type reports, valid at 0000, 0600, 1200, and 1800 UTC, were classified into one of either three or five mutually exclusive categories. The first classification follows the MOS precipitation type categories shown e.g. in Table 1 by Shafer (2015), and distinguishes 'freezing', 'frozen', and 'liquid' precipitation, classifying sleet as 'freezing' and any mixture of liquid precipitation with snow as 'liquid' (Allen and Erickson 2001a,b; Shafer 2015). This three category classification permits a direct comparison with the MOS technique used operationally by the Meteorological Development Laboratory of the National Weather Service (NWS). In addition, we consider a five-category classification which differs from the previous one in that it splits the 'freezing' category up into freezing rain (FZRA), freezing drizzle (FZDZ), and ice pellets (IP), in order to study in how far our forecasts are able to provide probabilistic guidance that reflects the new certification standards of the federal aviation administration (FAA) allowing some aircraft to fly in FZDZ but not FZRA. The frozen and liquid types are relabeled as snow (SN) and rain (RA) respectively. The observation data set used here clearly is not optimal for performing a five-category classification. Only a subset of the 551 stations are augmented by human observers and are able to report IP and FZDZ; all other stations may erroneously report a different type and thus contaminate both training and verification samples with T_w profiles that should be associated with IP or FZDZ but aren't. We accept the detrimental effect that this might have on the performance of our method because our priority is a direct comparison with Shafer (2015), but we note that our method might demonstrate somewhat better skill in distinguishing the different freezing precipitation types if it were trained with an observation data set like the one from the mPING project (Elmore et al. 2014) that is more consistent in how it reports IP, FZRA, and FZDZ.

b. Forecasts

Predictors used in this study were derived from the second-generation Global Ensemble Forecast System (GEFS) reforecast data set (Hamill et al. 2013). GEFS data were extracted for 2-m temperature and 2-m specific humidity on GEFS's native Gaussian grid at ~ 0.5 degree resolution in an area surrounding the CONUS, Alaska, and southern Canada over the period 1996-2013. Surface pressure and temperatures, specific humidities, and geopotential heights at the pressure levels 1000, 925, 850, 700, 500, and 300 hPa were obtained on a ~ 1 degree resolution grid covering the same area. Horizontal grids were bilinearly interpolated to the station locations and were used to convert the temperatures at the surface and at the pressure levels into wetbulb temperatures. Using the geopotential height fields, the wetbulb tempertures at each pressure level were associated with a certain height above ground level (AGL) where ground level here refers to the GEFS model grid. Vertical wetbulb temperature profiles were obtained by linear interpolation between those pressure levels, and then discrete values were taken at fixed heights up to 3000 m AGL. As a result of the rather coarse model grid resolution of ~ 0.5 degrees, the model grid elevation and the true elevation at the station locations differ substantially in regions with complex terrain. In order to adjust the resulting biases of the vertical wetbulb temperature profiles we:

- calculate the biases at the surface as the annual average difference between observed and (horizontally interpolated) analyzed wetbulb temperatures at each location;

- assume that the bias linearly decreases to zero at the 500 hPa level;

- correct the entire vertical wetbulb temperature profile accordingly, i.e. apply the full (additive) bias correction at the surface and gradually reduce the correction to zero with increasing height above the ground.

This procedure is *not* meant to correct complex forecast biases which potentially have a seasonal and diurnal cycle; these are somewhat implicitly addressed by the classification method described in the subsequent section. The procedure described here only tries to remove biases resulting from the mismatch between the terrain as represented by the forecast model and the true terrain.

For the remainder of this paper it is our working assumption that the observed precipitation type only depends on the vertical wetbulb temperature profile above the ground, i.e. given two identical profiles the outcome is independent of the location and time of the year at which those profiles were observed. This is a simplification that does not account for the microphysical forcing (precipitation rate, degree of riming, etc.) but is necessary as it allows us to pool data across all stations and across all dates within the cool seasons considered here. Fig. 1 depicts examples of wetbulb temperature profiles at initialization time (i.e. based on GFS analyses) obtained as described above. There are typically only three or four pressure levels which data are used for the reconstruction of the section of the profiles shown in this figure. It is clear that the resulting interpolation error (compare e.g. with Fig. 2 in Reeves et al. 2014 who study wetbulb temperature profiles obtained from radiosonde observations) can be substantial, and adds to the overall uncertainty about the vertical profiles resulting from initial condition and forecast uncertainty.

3. Regularized Bayesian Classification

The method proposed in this paper is based on Bayes' Theorem, which has recently been employed by Hodyss et al. (2016) to derive optimal weights for different model forecasts and climatology in a statistical postprocessing approach for continuous predictands. In our setting, the predictand is categorical, but the same general principle can be used as a starting point. Assume that we know, for each location s and each date t (day of the year and time of the day), the climatological probability for each precipitation type $k \in \{1, ..., K\}$ to occur. Denote this probability

by π_{kst}. Assume further that for each k we know the multivariate probability density function (PDF) φ_k that characterizes the distribution of the discretized, predicted vertical wetbulb temperature profiles that are compatible with the observed precipitation type k. For efficient statistical classification, we want these distributions to be as different as possible. Fig. 1 suggests that the differences between profiles corresponding to the different precipitation types are much more pronounced in the lower half of the sections depicted in the plots, and we therefore only consider wetbulb temperatures corresponding to heights above the surface up to 1500 m, even though the precipitation-generation layer is usually far outside this range. Sampling the profiles every 100 m then leaves us with a wetbulb temperature vector of dimension $d = 16$, and φ_k models the probability distribution of this vector for each k. Due to our assumption that given two identical profiles the observed precipitation type should not depend on s and t, φ_k is assumed constant across the entire spatial domain and throughout the year (it may vary with lead time though since there is typically more dispersion around the mean profile for longer leads). According to Bayes' Theorem (Wilks 2006, Eq. (13.32)), given the climatological probabilities π_{kst}, the PDFs φ_k, and a new, predicted vector $\boldsymbol{x}$ of vertical wetbulb temperature profile values, the conditional probability $P(k|\boldsymbol{x})$ of observing precipitation type k is

$$P(k|\boldsymbol{x}) = \frac{\pi_{kst}\,\varphi_k(\boldsymbol{x})}{\sum_{i=1}^{K} \pi_{ist}\,\varphi_i(\boldsymbol{x})}. \tag{1}$$

In this study we approximate the climatological probabilities π_{kst} by the relative frequencies of observed precipitation types, calculated separately for each location s, each month (but pooling all days within a month and all years for which data are available), and each time of the day. The following subsections discuss how an adequate model for φ_k can be defined and fitted. Note that Eq. (1) is also the starting point for *(quadratic) discriminant analysis*, where a deterministic classification rule is derived from this equation.

a. Basic model: multivariate normal distribution

A standard assumption with this approach to probabilistic classification is to let be φ_k a multivariate normal PDF (Wilks 2006, Sec. 13.3.3). This PDF is completely characterized by its mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. Given a set of training data we can estimate $\boldsymbol{\mu}_k$ as the empirical mean and $\boldsymbol{\Sigma}_k$ as the empirical covariance matrix of the subset of the training profiles that correspond to an observed precipitation type k. In order to focus on situations where the outcome is truly uncertain, we only use locations/dates for the calculation of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ for which $\pi_{kst} < 0.99$ for every $k \in \{1, ..., K\}$. This excludes, for example, precipitation events in January at high altitudes where precipitation most likely occurs in the form of snow, and precipitation events in May in Florida, where precipitation almost surely occurs in the form of rain. These events will still be used for validation, but excluding them for estimating the PDFs φ_k moves the mean vectors for rain and snow closer to the freezing point and improves the distribution fit in this temperature range where classification is most challenging. This results in a noticeable improvement in probabilistic classification skill, and one could even try to optimize the probability threshold for omitting cases from the training data set. However, we do not expect much further improvement from lowering the threshold and keep it fixed at 0.99.

Given $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, and a wetbulb temperature vector $\boldsymbol{x}$, the likelihood $\varphi_k(\boldsymbol{x})$ in Eq. (1) under the assumption of a multivariate normal distribution is given by

$$\varphi_k(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \, |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)} \tag{2}$$

where $\boldsymbol{x}'$ denotes the transpose of $\boldsymbol{x}$. Using the eigenvalue decomposition $\boldsymbol{\Sigma}_k = \boldsymbol{E}_k \boldsymbol{\Lambda}_k \boldsymbol{E}_k'$ to transform $\boldsymbol{x}$ into vectors $\boldsymbol{u}_k = \boldsymbol{E}_k'(\boldsymbol{x} - \boldsymbol{\mu}_k)$ of centered principal components (PCs), this likelihood

can be expressed as a product of univariate likelihoods

$$\varphi_k(\boldsymbol{x}) = \prod_{j=1}^{d} \phi_{0,\lambda_{k,j}}(u_{k,j}) \,, \tag{3}$$

where $\phi_{(0,\lambda_{k,j})}$ is the PDF of a univariate normal distribution with mean 0 and variance equal to the j-th eigenvalue $\lambda_{k,j}$ of $\boldsymbol{\Sigma}_k$. This re-interpretation of $\varphi_k(\boldsymbol{x})$ in terms of principal components will later be used to motivate an easily interpretable and computationally efficient generalization of the basic multivariate normal model discussed above. Fig. 2 shows the means of the $K = 5$ classes of interest and the variability around the respective mean in the direction of the first eigenvector of $\boldsymbol{\Sigma}_k$. The different shapes of these eigenvectors suggest that there is structural information in the wetbulb temperature profiles beyond the mean that can be utilized for classification.

b. First extension: introducing skewness

Upon closer inspection, the assumption of a multivariate Gaussian distribution made above turns out to be a coarse approximation of the truth. For example, wetbulb temperature profiles much cooler than the mean profile are still compatible with observing snow, whereas the probability for observing snow but predicting a relatively warm profile decreases more rapidly (such profile would more be associated with observing rain). Applying a power transformation to each component of the wetbulb temperature vectors can make the distributions more symmetric (Wilks 2006, Sec. 3.4.1), but their direct physical interpretation is lost in that process. Alternatively, a more complex, multivariate skew normal distribution could be used to fit the untransformed data (Azzalini and Capitanio 1999). In our setting with dimension $d = 16$, this requires estimating a large number of model parameters which is computationally and numerically challenging. Here, we propose a similar approach that permits an intuitive interpretation and straightforward statistical inference. We first proceed as described above, estimate $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ as the as the empirical means

and covariance matrices of the wetbulb temperature vectors associated with each precipitation type, and use them to calculate the centered PCs $u_{k,1},\dots,u_{k,d}$ of each wetbulb temperature vector $\boldsymbol{x}$. Possible skewness can then be addressed for each PC separately by modeling them by univariate skew normal distributions $f_{\xi_{k,j},\omega_{k,j},\alpha_{k,j}}$ with location parameter $\xi_{k,j}$, scale parameter $\omega_{k,j}$, and shape parameter $\alpha_{k,j}$. By construction, the PCs are centered and have variances $\lambda_{k,j}$, so for given $\alpha_{k,j}$ the location and scale parameters are determined by

$$\omega_{k,j}^2 = \lambda_{k,j}\left(1-\frac{2\alpha_{k,j}^2}{\pi(1+\alpha_{k,j}^2)}\right)^{-1} \quad \text{and} \quad \xi_{k,j} = -\omega_{k,j}\sqrt{\frac{2\alpha_{k,j}^2}{\pi(1+\alpha_{k,j}^2)}}. \tag{4}$$

Using these relations, $\alpha_{k,j}$ can be estimated via maximum likelihood. Since the impact of the PCs for smaller eigenvalues on classification will be de-emphasized as explained in the next subsection, we only bother to estimate $\alpha_{k,j}$ for $j\in\{1,2\}$, and set $\alpha_{k,j}=0$ (i.e. no skewness) for all other PCs. Fig. 3 shows histograms of the first three PCs and the fitted distributions. The variability in the direction of the first eigenvector corresponds to cooler/warmer than average wetbulb temperatures of the entire vertical profile (see Fig. 2), and the asymmetry of the associated PCs as described above for snow is clearly visible in those histograms. The fitted skew normal distributions are capable of modeling this asymmetry, and for calculating the likelihood $\varphi_k(\boldsymbol{x})$ one only needs to replace Eq. (3) by

$$\varphi_k(\boldsymbol{x}) = \prod_{j=1}^{d} f_{\xi_{k,j},\omega_{k,j},\alpha_{k,j}}(u_{k,j})\,. \tag{5}$$

c. Second extension: regularization

A further modification to the multivariate PDF φ_k is required to make this Bayesian classification method work efficiently. As pointed out in Sec. 2, the reconstruction of the vertical wetbulb temperature profile based on the GEFS model output at a few available pressure levels comes with substantial interpolation errors, and especially features at small vertical scales are not resolved.

On the other hand, even if we could reconstruct those profiles at high vertical resolution, it is unclear whether their fine-scale structure carries any useful information for the discrimination between different precipitation types. In the light of the principal component interpretation of the multivariate likelihood $\varphi_k(\boldsymbol{x})$ discussed above, it would seem natural to truncate after a few PCs and omit the last few terms in the product in Eq. (5), which typically correspond to eigenvectors representing the fine-scale structure of the vertical profiles. This is problematic, however, since the different precipitation types can have very different leading eigenvectors (see Fig. 2) and different spectra of eigenvalues, and so both the fractions of explained variances and the subspaces onto which the profiles are projected would be different. A more appropriate way to mute the effect of higher PCs on the likelihood $\varphi_k(\boldsymbol{x})$ is to *regularize* the covariance matrix $\boldsymbol{\Sigma}_k$, i.e. to replace it by

$$\tilde{\boldsymbol{\Sigma}}_k(a_k, b_k) := a_k \boldsymbol{\Sigma}_k + b_k \mathrm{I}, \tag{6}$$

where I is the identity matrix and a_k, b_k are positive coefficients for which selection will be discussed later. This idea of regularization was introduced by Friedman (1989) in the context of a similar but deterministic classification technique referred to as *regularized discriminant analysis*. In our probabilistic setting we will refer to this idea as *regularized Bayesian classification (RBC)*. It can easily be combined with our assumption of skew normal distributions of the PCs by noting that the regularization in Eq. (6) leaves the eigenvectors unchanged but turns the eigenvalues $\lambda_{k,j}$ into

$$\tilde{\lambda}_{k,j} = a_k \lambda_{k,j} + b_k, \qquad j = 1, \ldots, d\,. \tag{7}$$

Setting $a_k := 1 - b_k/\lambda_{k,1}$ leaves the first eigenvalue $\lambda_{k,1}$ unchanged but increases all other eigenvalue with the relative increase being larger for smaller eigenvalues. This implies an artificial inflation of the variances of the univariate, skew normal PDFs in Eq. (5), and causes the corresponding likelihoods to be relatively less sensitive to the PCs $u_{k,j}$ corresponding to the smaller eigenvalues.

To find the optimal degree of inflation, i.e. optimal regularization parameters $b_1, \ldots, b_K$, we use the training data set (forecasts and observations) that was used to estimate $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, and proceed as follows:

- for given parameters $b_1, \ldots, b_K$, and for every wetbulb temperature profile $\boldsymbol{x}$ in the training data set, use Eqs. (1), (5), (4), and (7) to calculate the likelihoods $\varphi_k(\boldsymbol{x})$ and resulting forecast probabilities $P(k|\boldsymbol{x})$ for each k; and

- use the corresponding training observations to calculate the resulting Brier skill scores BSS_k (see Sec. 4 for a definition) and choose $b_1, \ldots, b_K$ such that the sum $\sum_{k=1}^{K} \mathrm{BSS}_k$ is maximized.

Note that the target function $\sum_{k=1}^{K} \mathrm{BSS}_k$ that we seek to maximize gives the same weight to all precipitation type categories despite their very different frequencies of occurrence. This is done on purpose to foster good performance of our method with regard to the rare freezing precipitation types. Different priorities can be set, however, by introducing weights that increase or decrease the impact of the skill for certain precipitation types on the target function. In our example, the optimal values of b_k were between 5 and 10 for all k, which is smaller than the second, but about 2-3 times larger than the third eigenvalues of $\boldsymbol{\Sigma}_k$ (see Fig. 3). This suggests that useful information about the vertical structure of the wetbulb temperature profiles is limited to the first two principal components.

d. Third extension: applying RBC to ensemble forecasts

The RBC approach presented above yields forecast probabilities for the occurrence of each precipitation type given a single (i.e. deterministic) forecast of a vertical wetbulb temperature profile. In our situation where we have an ensemble of forecasts, this ensemble represents some

of the uncertainty about the predicted vertical wetbulb temperature profiles, and its use can thus reduce the amount of variability that is modeled purely statistically.

We proceed as before regarding the estimation of $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, and the skewness parameters $\alpha_{k,j}$, considering the forecast profiles $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ of the M ensemble members as separate cases. Centering and projecting those profiles onto the eigenvectors of $\boldsymbol{\Sigma}_k$ yields principal components $u_{k,j,m}$ and M different likelihoods $\varphi_k(\boldsymbol{x}_m)$ for each class. The resulting probability forecasts $P(k|\boldsymbol{x}_m)$ can be combined to a single probability forecast by simply taking the mean for each class

$$P(k|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M) = \frac{1}{M} \sum_{m=1}^{M} P(k|\boldsymbol{x}_m) \tag{8}$$

This way of linear pooling, however, does in general not yield reliable probability forecasts even if all of the individual member probability forecasts $P(k|\boldsymbol{x}_m)$ are reliable (Ranjan and Gneiting 2010). Indeed, as pointed out above, the simultaneous consideration of different ensemble member forecasts explains some of the variability of the forecast profiles, and so in return the statistically modeled variability needs to be reduced to avoid under-confident probability forecasts. We do this by replacing the variances $\lambda_{k,j}$ of the principal component PDFs that were originally obtained as the eigenvalues of $\boldsymbol{\Sigma}_k$ by the empirical variances $v_{k,j} = \mathrm{var}(\bar{u}_{k,j})$ of the ensemble-mean PCs

$$\bar{u}_{k,j} = \frac{1}{M} \sum_{m=1}^{M} u_{k,j,m} \,.$$

While this is equivalent to just operating on the ensemble-mean profiles, evaluating Eq. (1) with an ensemble-mean profile does not yield the same probabilities as Eq. (8) due to the non-linearity of the likelihood function; Eq. (8) averages the probabilities corresponding to the different atmospheric situations represented by the ensemble, as opposed to averaging the vertical profiles and deriving a probability from the averaged state of the atmosphere. In contrast to $\lambda_{k,j}$, which describes the variability of a certain PC over all profiles in the training data set corresponding to a certain precipitation type, $v_{k,j}$ elides the variability within the ensemble, and is therefore smaller

than $\lambda_{k,j}$. All subsequent steps, i.e. regularization according to Eq. (7), calculation of likelihoods according to Eqs. (5), (4), and calculation of probability forecasts according to Eqs. (1), (8), remain the same as before, but are carried out based on $v_{k,j}$ instead of $\lambda_{k,j}$. The results in the following section will show that this reduction of statistically modeled variability in favor of dynamically explained variability yields a noticeable improvement of predictive performance at longer forecast lead times.

4. Results

To test our RBC approach and compare it against the operational MOS technique, we adopt the verification setup used by Shafer (2015), but studying only the case of a training sample comprised of five cool seasons (see Sec. 2). Forecasts were produced and verified for the cool seasons 2001 through 2012. For each of these verification seasons, the methods were trained with data from the previous five cool seasons, i.e. the statistical model used for producing probability forecasts for the cool season 2001 was set up based on data (forecasts and observations) from the cool seasons 1996 through 2000. For estimating the climatological frequencies, which are used as a reference forecast on the one hand, and as a prior distribution for our RBC technique on the other hand, the entire observation record was used. This may be justified by noting that in practice long time series of observations are often available while forecast systems keep evolving and available forecast time series from a stable system are typically much shorter.

First, we assess the reliability of the RBC probability forecasts for different lead times separately for each precipitation type. Fig. 4 depicts reliability diagrams for probability forecasts generated by the ensemble-based version of the RBC method. For all forecast lead times considered in this study (including those not shown in this figure), the curves are close to the diagonal, which means

that the relative frequency of occurrence of each precipitation type matches the probability with which it was predicted.

The RBC probability forecasts based on the GEFS control run only were equally reliable (not shown here), so as a second validation tool we consider a quantitative performance measure, the Brier skill score (Wilks 2006, Eqs. 7.34 and 7.35). In addition to reliability, the Brier score evaluates the resolution of a forecast, i.e. its ability to distinguish situations with different frequencies of occurrence. A skill score relates the score of the forecast method of interest to a reference score (here: climatological frequency of occurrence, calculated separately for each location, each month, and each time of the day) and thus facilitates its interpretation (Wilks 2006, Sec. 7.33). Here, we compare the Brier skill scores (BSSs) of the generalized operator equation (GOE) implementation of the operational MOS PoPT technique described in (Shafer 2010, 2015) and GEFS ensemble mean forecasts, the RBC method based on the GEFS control run only, and the RBC method using each of the individual ensemble member forecasts for forecast lead times up to 192h. The MOS PoPT approach currently only distinguishes three classes: frozen (SN), liquid (RA), and freezing (IP, FZRA or FZDZ). To allow a direct comparison, we aggregate the five class RBC probabilities to three class probabilities, and compare the BSSs for the three class probabilities of all three methods on the one hand, and the BSSs for the IP, FZRA and FZDZ probabilities by the two RBC implementations on the other hand. The results depicted in Fig. 5 permit several conclusions:

- the use of ensemble forecasts as opposed to a single deterministic run clearly benefits forecast performance, especially for longer forecast lead times;

- for the frozen and liquid class, the improvement of the RBC ensemble method over the MOS PoPT approach is marginal; if the MOS PoPT approach were extended such as to use the

individual ensemble member forecasts rather than the ensemble mean, there may be no improvement at all;

- for the particularly challenging, freezing category, however, there is a noticeable benefit of using a statistical method (such as RBC) that can use the full vertical wetbulb temperature profile as a predictor; and

- the skill for the freezing categories (especially IP and FZDZ) is low compared to the skill for RA and SN; yet our RBC method can provide skillful probabilistic guidance on freezing precipitation several days ahead, and even has the potential to separate IP, FZRA and FZDZ.

The results discussed above show the effectiveness of our RBC method in general and the utility of ensemble forecasts in particular. How about the other two extensions (modeling skewness of the PCs, regularization), how much do they contribute to the skill of the RBC approach? How much skill is lost if the available training data for estimating $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, b_k, and $\alpha_{k,j}$ is composited of just one instead of five cool seasons? To answer these questions we use the control run based RBC method (fitted with five years of training data, as above) as a benchmark and compare it to a) the same model fitted with training data from a single cool season, b) a simplified model that regularizes $\boldsymbol{\Sigma}_k$ according to (6) but assumes normal instead of skew normal distributions of the PCs, c) a simplified model that uses skew normal distributions but does not regularize the empirical covariance matrices. The following conclusions can be drawn from the results shown in Fig. 6:

a) Reducing the training sample size hardly affects the performance in predicting SN and RA probabilities, but has a rather strong, negative impact on the predictive performance for IP, FZRA, and FZDZ. For the two former, there are still enough cases within a single cool season to warrant a good estimation of model parameters. Estimating the parameters for the rare,

freezing precipitation types, however, requires either several years of training data or a much denser observation network. In addition to the issue of boundary discontinuity, this is also an argument in favor of a pooling data across all locations as opposed to partitioning the country into more homogeneous sub-domains. The latter might better account for different regional characteristics, but Fig. 6 suggests that these benefits could be nullified by the concomitant reduction of training sample size.

b) Simplifying the RBC approach by assuming a multivariate normal distribution for the wetbulb temperature profiles affects the predictive performance in the opposite way. While the more flexible distribution model does not seem to benefit the freezing precipitation types, the better approximation of the distributions of SN and RA profiles that results from modeling skewness in the PCs translates into improved skill of the resulting probability forecasts.

c) Finally, Fig. 6 highlights the necessity of regularizing the empirical covariance matrices. Without regularization, skill drops dramatically for SN and RA and becomes negative beyond a forecast lead time of three days. For the freezing types the impact is even stronger and lack of regularization results in Brier skill scores around -1.0 for all lead times. Unregularized classification gives as much emphasis to the noisy, unwarranted fine scale structure of the wetbulb temperature profiles as it gives to the first PCs that represent meaningful features of these profiles, and this results in probability forecasts that are entirely off the mark.

To illustrate the capabilities and limits of probabilistic guidance obtained with the RBC method applied to GEFS ensemble forecasts, two particular cases studies presented. Fig. 7 shows spatial maps of FZRA probabilities for 27 January 2009, 0000 UTC, with a forecast lead time of 2, 4, and 6 days ahead. This date is in the middle of a major ice storm that impacted parts of Oklahoma, Arkansas, Missouri, Illinois, Indiana, West Virginia, and Kentucky. The plots suggest that

the GEFS captured the atmospheric situation well, and the RBC methods provides a strong probabilistic signal for freezing rain even at 6 days of lead time. For the event shown in Fig. 8 (also studied by Reeves et al. 2016) the situation is more complex. The plots show observed precipitation types and 2 day ahead RBC forecast probabilities for 22 February 2013, 0000 UTC. Even at this short lead time, the probabilistic signal for the freezing precipitation types is rather weak (note the different color scales) and no clear guidance is provided as to which particular freezing precipitation type will dominate in each geographical area. This underscores the inherent uncertainty in precipitation type forecasts based on a global ensemble prediction system, and illustrates the limits of such forecasts. Notwithstanding, the RBC probability forecasts indicate an increased risk of freezing precipitation, and we believe that there is substantial value in communicating that risk to decision makers.

5. Discussion

In this paper we have proposed a method for conditional probabilistic precipitation type forecasting which is based on a statistical model for the predicted vertical wetbulb temperature profiles that are compatible with each precipitation type. Using Bayes' theorem this model can be inverted such that it yields probability forecasts for each precipitation type given a new predicted profile.

There were many sources of forecast and data uncertainty that needed to be accounted for in a precipitation typing methodology. These include forecast errors stemming from initial condition uncertainty, from model error, and in this case from the need to interpolate NWP model output from a relatively coarse horizontal grid and a few pressure levels to a much finer horizontal and vertical resolution and more complex orography at the surface level. Availability of sigma-level forecast data at a finer vertical resolution could reduce this last component of uncertainty, which contributes noticeably to the overall uncertainty about the wetbulb temperature profiles at short

lead times. It is suggested that thermodynamic variables be archived at many vertical levels above the surface when generating future reforecasts. At longer lead times, forecast errors become the dominant source of uncertainty, and the interpolation error might be negligible. At short lead times, forecasts from a high resolution, limited-area NWP model might be available, which might be accurate enough to yield superior classification results using an explicit precipitation type diagnosis scheme (e.g. Benjamin et al. 2016) or the spectral bin classifier proposed by Reeves et al. (2016), but such guidance could be leveraged in a probabilistic framework, too.

The strength of the method proposed here is that it can handle the large uncertainty that inevitably comes with predictions from a global forecast system, and that it can still provide reliable, probabilistic precipitation type forecasts at forecast lead times up to seven days ahead. It has sufficient skill to give decision makers at least a heads up about precipitation type related weather risks, and it can easily be extended to distinguish further precipitation type classes like mixtures of snow and rain, mixtures of freezing precipitation types, and so forth, if they are reported accurately in the observations. The observation data set used here is not optimal in that regard as it is inconsistent in how it reports IP, FZRA, and FZDZ, and the skill of our method in distinguishing these types might actually be better than reported here if it were trained with an observation data set like the one from the mPING project (Elmore et al. 2014, 2015) in which IP, FZRA, and FZDZ are distinguished more systematically.

We have focused on vertical profiles of wetbulb temperature as a predictor variable. However, by combining the statistical dimension reduction / regularization techniques used here with more physically motivated aggregation methods one might be able to further improve skill by using additional predictors such as relative humidity profiles. Alternatively, one could use modern machine learning techniques like neural networks to identify features of vertical wetbulb temperature and humidity profiles that determine the observed precipitation type. While extremely powerful, these

techniques typically require large data sets for training, but these may become available once several years of mPING data have been collected, and allow one to explore the more data-intensive machine learning techniques.

Acknowledgments. The authors thank Amanda Hering for useful discussions which inspired some of the refinements to our basic distribution model for the wetbulb temperature profiles. Our research was supported by grants from the NOAA/NWS Sandy Supplemental (Disaster Relief Appropriations Act of 2013) and the NOAA/NWS Research to Operations (R2O) initiative for the Next-Generation Global Prediction System (NGGPS), award # NA15OAR4320137.

References

Allen, R. L., and M. C. Erickson, 2001a: AVN-based MOS precipitation type guidance for the United States. NWS Technical Procedures Bulletin No. 476, NOAA, U.S. Dept. of Commerce, 8 pp.

Allen, R. L., and M. C. Erickson, 2001b: MRF-based MOS precipitation type guidance for the United States. NWS Technical Procedures Bulletin No. 485, NOAA, U.S. Dept. of Commerce, 8 pp.

Azzalini, A., and A. Capitanio, 1999: Statistical applications of the multivariate skew normal distribution. *J. Roy. Stat. Soc. B*, **61**, 579–602.

Baldwin, M., R. Treadon, and S. Contorno, 1994: Precipitation type prediction using a decision tree approach with NMCs mesoscale eta model. *10th Conf. on Numerical Weather Prediction*, Portland, OR, Amer. Meteor. Soc., 30–31.

Benjamin, S. G., J. M. Brown, and T. G. Smirnova, 2016: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud-precipitation microphysics parameterization. *Wea. Forecasting*, **31**, 609–619.

Bourgouin, P., 2000: A method to determine precipitation type. *Wea. Forecasting*, **15**, 583–592.

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, H. D. Reeves, and L. S. Rothfusz, 2014: mPING: Crowd-sourcing weather reports for research. *Bull. Amer. Meteor. Soc.*, **95**, 1335–1342.

Elmore, K. L., H. M. Grams, D. Apps, and H. D. Reeves, 2015: Verifying forecast precipitation type with mPING. *Wea. Forecasting*, **30**, 656–667.

Friedman, J. H., 1989: Regularized discriminant analysis. *J. Amer. Stat. Assoc.*, **84**, 165–175.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. G. Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast data set. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565.

Hodyss, D., E. Satterfield, J. McLay, T. M. Hamill, and M. Scheuerer, 2016: Inaccuracies with multimodel postprocessing methods involving weighted, regression-corrected forecasts. *Mon. Wea. Rev.*, **144**, 1649–1668.

Ramer, J., 1993: An empirical technique for diagnosing precipitation type from model output. *Fifth Int. Conf. on Aviation Weather Systems*, Vienna, VA, Amer. Meteor. Soc., 227–230.

Ranjan, R., and T. Gneiting, 2010: Combining probability forecasts. *J. Roy. Stat. Soc. B*, **32**, 71–91.

Reeves, H. D., K. L. Elmore, A. V. Ryzhkof, T. J. Schuur, and J. Krause, 2014: Sources of uncertainty in precipitation-type forecasting. *Wea. Forecasting*, **29**, 936–953.

Reeves, H. D., A. V. Ryzhkof, and J. Krause, 2016: Discrimination between winter precipitation types based on spectral-bin microphysical modeling. *J. Appl. Meteor. Climatol.*, **55**, 1747–1761.

Schuur, T. J., H.-S. Park, A. V. Ryzhkof, and H. D. Reeves, 2012: Classification of precipitation types during transitional winter weather using the ruc model and polarimetric radar retrievals. *J. Appl. Meteor. Climatol.*, **51**, 763–779.

Shafer, P., 2010: Logit transforms in forecasting precipitation type. *20th Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., P222.

Shafer, P. E., 2015: A sample size sensitivity test for MOS precipitation type. *Special Symposium on Model Postprocessing and Downscaling*, Phoenix, AZ, Amer. Meteor. Soc., 4.1.

Stewart, R. E., J. M. Thériault, and W. Henson, 2015: On the characteristics of and processes producing winter precipitation types near 0°C. *Bull. Amer. Meteor. Soc.*, **96**, 623–639.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, International Geophysics Series, Vol. 91. 2nd ed., Elsevier Academic Press.

LIST OF FIGURES

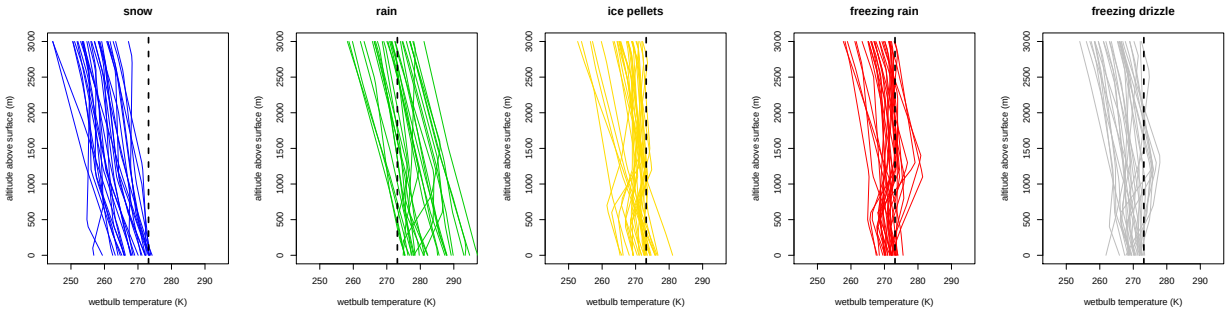


FIG. 1. Approximate vertical wetbulb temperature profiles reconstructed from GEFS forecasts fields at initialization time. For each of the five precipitation types of interest, 30 profiles are depicted that were randomly sampled from locations/dates where that precipitation type was reported.

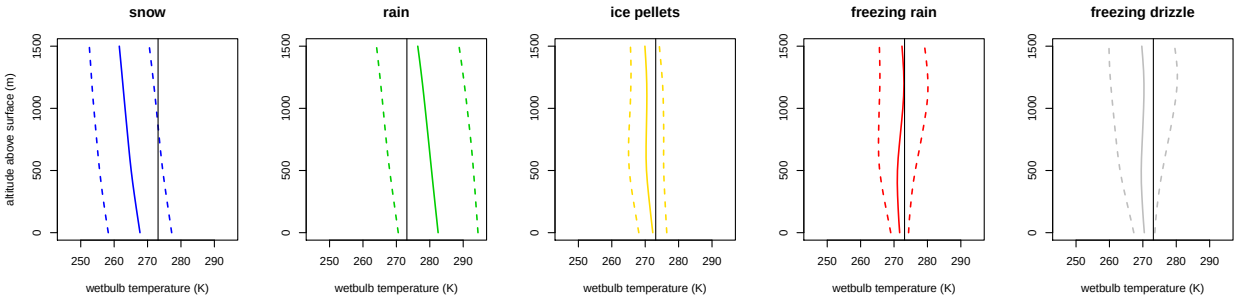
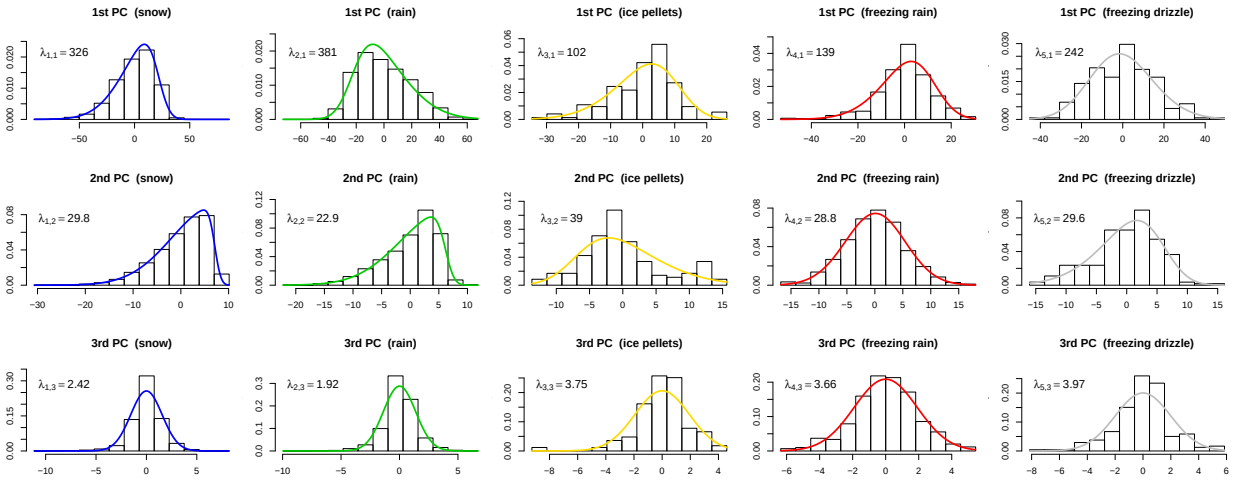
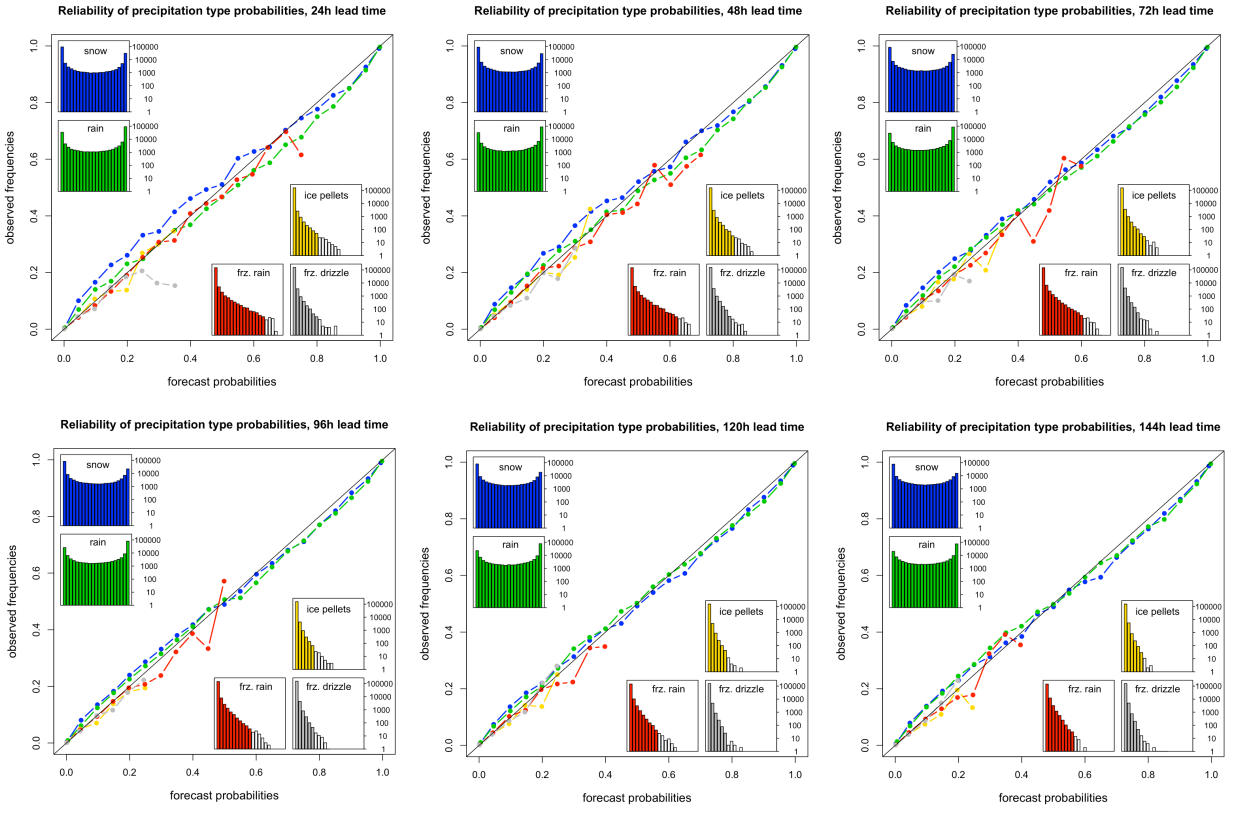


FIG. 2. Empirical means (solid lines) and variability (two standard deviations, dashed lines) in the direction of the first eigenvector of $\mathbf{\Sigma}_k$ for each of the five precipitation types distinguished by our algorithm.



511 FIG. 3. Histograms and fitted skew normal distributions for the first three principal components of the wetbulb
512 temperature profile vectors of each class.



513 FIG. 4. Reliability diagrams for probability forecasts at various forecast lead times generated by the ensemble-
514 based version of the RBC method with probabilities rounded to a precision of 0.05. The inset histograms depict
515 the frequency (on a logarithmic scale) with which the respective probabilities were forecast (x-axes are the same
516 as in the reliability diagrams). Points of the reliability curve associated with very infrequent forecast probabilities
517 (< 25 cases) are subject to substantial sampling variability and have therefore been omitted.

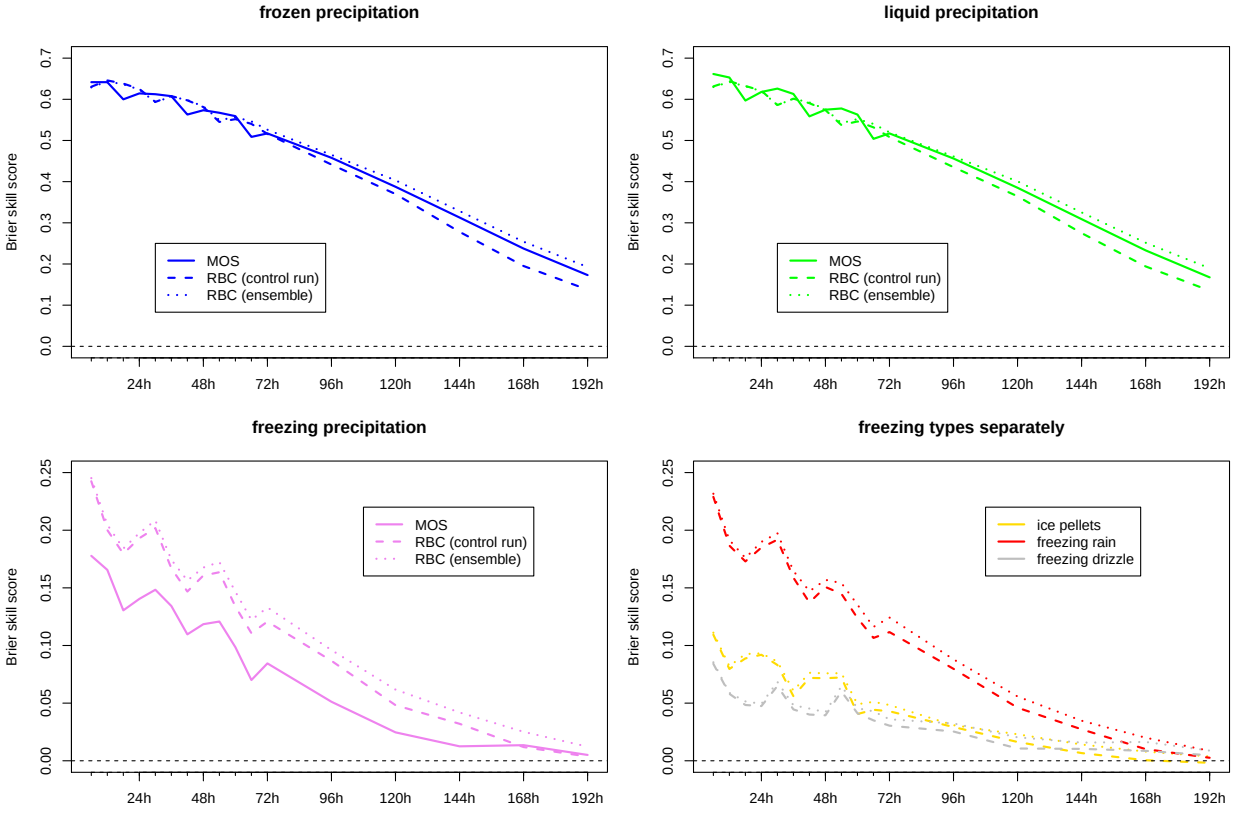


FIG. 5. Brier skill scores of the probability forecasts obtained with the operational MOS approach and the RBC methods based on the control run and based on the full ensemble.

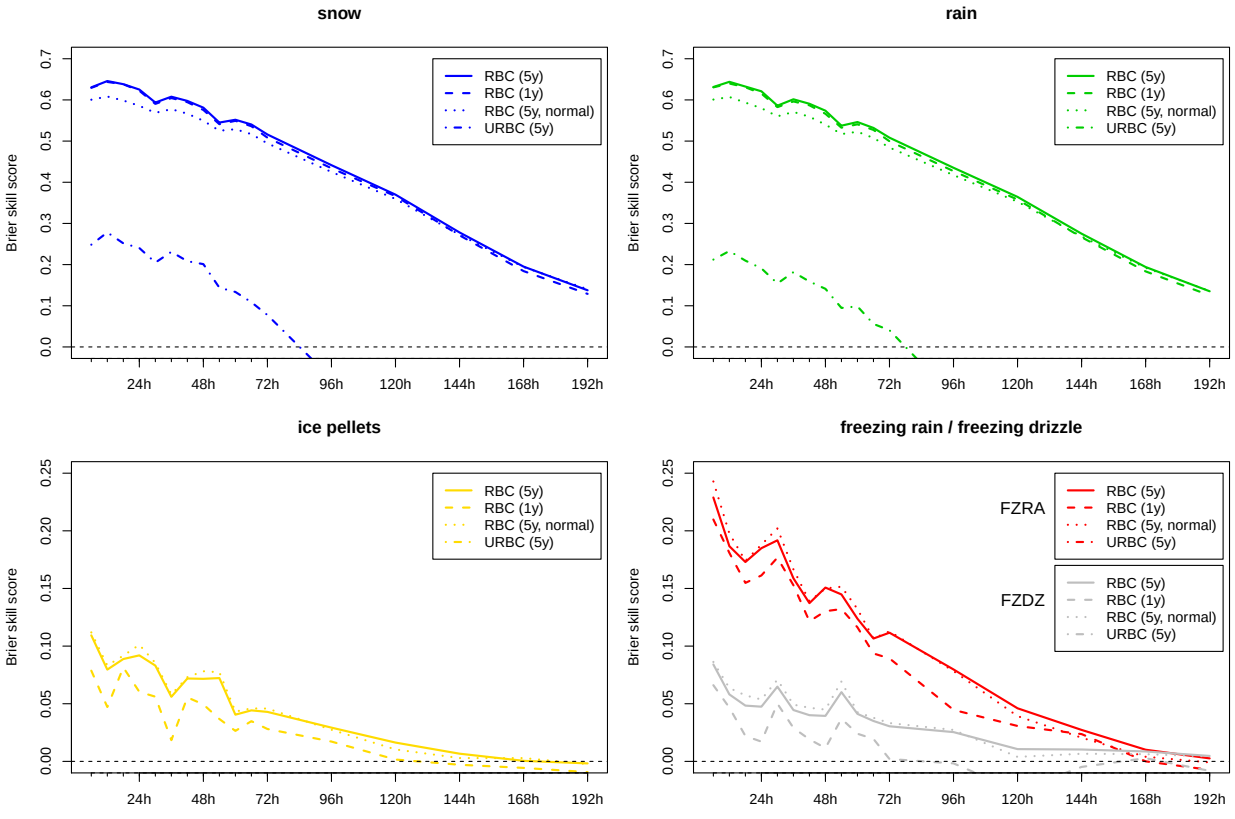


FIG. 6. Brier skill scores of the probability forecasts obtained with variants of the RBC method (applied to the GEFS control forecast) which use only one year of training data, normal instead of skew normal distributions, or unregularized covariance matrices $\mathbf{\Sigma}_k$.

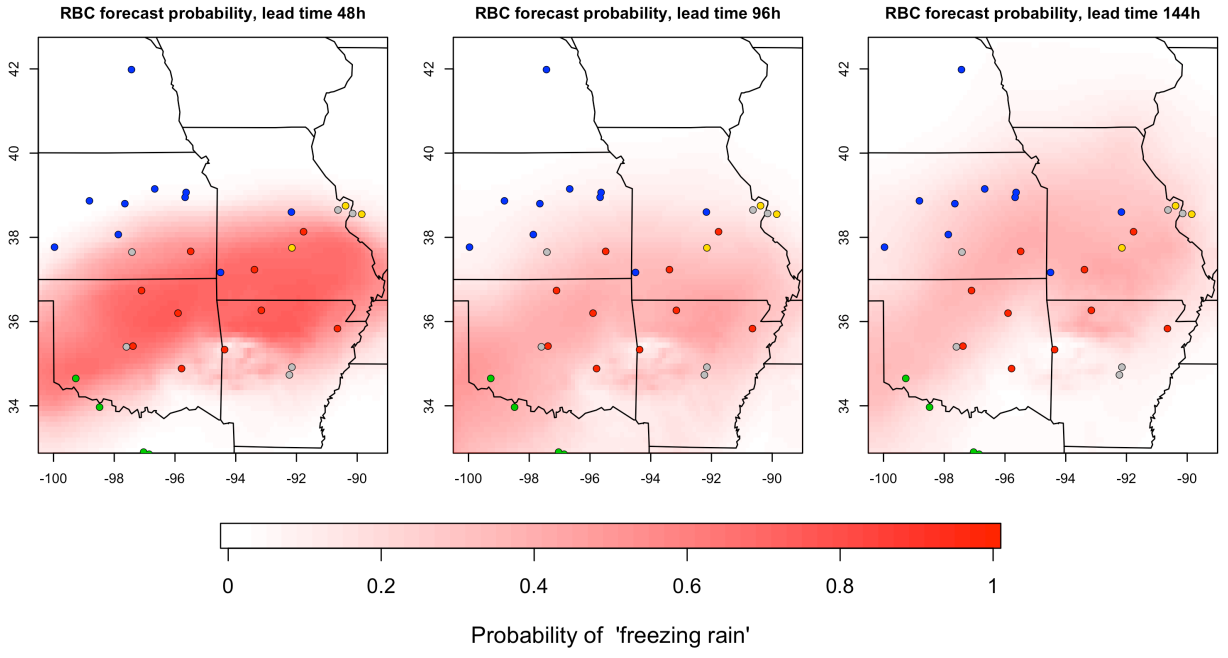


FIG. 7. Observed precipitation types (shaded circles with color scheme as in previous figures) on 27 January 2009, 0000 UTC, and predicted freezing rain probabilities by the ensemble-based RBC method for different forecast lead times.

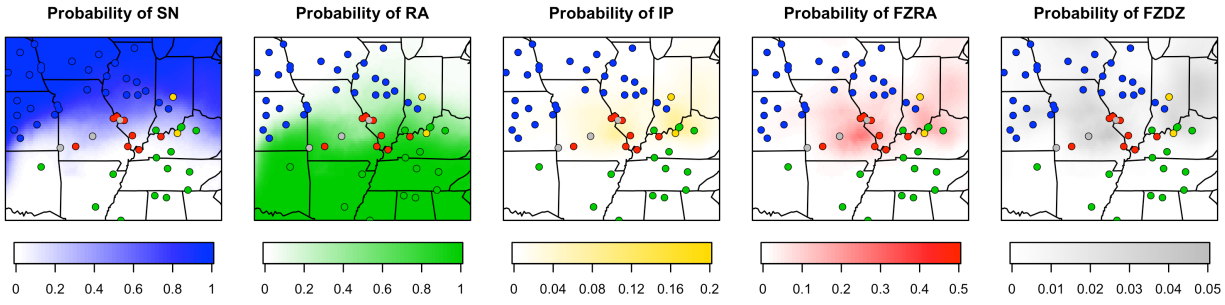


FIG. 8. Observed precipitation types on 22 February 2013, 0000 UTC, and predicted precipitation type probabilities by the ensemble-based RBC method with a forecast lead time of 48h.